Science, Vol 292, Issue 5525, 2264-2266 , 22 June 2001
[DOI: 10.1126/science.1059393]

◄ **Previous** Article ● Table of Contents ● **Next** Article ►

GENOMICS:

On Choosing Mammalian Genomes for Sequencing

Stephen J. O'Brien, Eduardo Eizirik, William J. Murphy*

Fifty years ago, James Rettie proposed a graphic imagery (1) that made the geological time frame of biological evolution more comprehensible. Rettie imagined a time-lapse motion picture of Earth taken from space, beginning 757 million years ago, with one image being photographed each year. Projected at the normal speed of 24 images per second, the resulting movie would take a year to view, with each day representing 2.1 million years. Here is what the movie would show:

From January to March there is little sign of life, then the first unicellular microbes appear in early April, giving rise to small multicellular aggregates later that month. In May the vertebrates emerge, and by July land plants have begun to cover the globe. In mid-September early reptiles preview the dawn of the dinosaur era, which continues through late November, dominating the world for 70 days. Birds and small mammals first appear in early November but are overshadowed by reptilian species until 1 December, when the dinosaurs disappear abruptly. By late December the recognizable ancestors of modern families of mammals make their debut, but not until midday on New Year's Eve do our first ancestors appear. Between 9:30 and 10 p.m., *Homo sapiens* migrates out of Africa to populate Eurasia and the Americas. At 11:54 p.m., recorded human history and civilization as we know it begin. Mammals flourish for the last 50 to 60 days of the movie year, and humankind eventually appears during the final 12 hours of the last day of the year.

Today, some 4600 to 4800 species of mammals dominate the planet. They occupy every continent and diverse ecological niches. The morphological and physiological differentiation seen among mammals is enormous, ranging from blue whales to echolocation-driven bats, from blind subterranean naked mole rats to us. The richness of mammalian species diversity and their remarkable adaptations have provided the evolutionary framework from which the human species evolved. The genomes of mammalian species encode the script for individual developmental distinctions, as well as the relict sequence records of historical events through which the genomes (and the organisms they prescribe) were sculpted by natural selection.

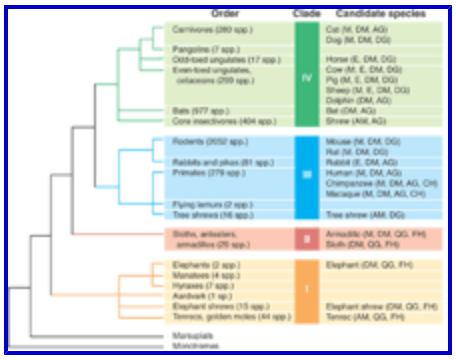
With the unveiling of draft versions of the human genome earlier this year, resources are now being harnessed to annotate the 30,000-odd human genes that direct our development, appearance, behavior, talents, and susceptibility to disease (2, 3). The interpretation of human genome organization will draw, to a large extent, from the genomes of other organisms, helping us to infer the function, regulation, and origins of our own genes (4-11). The value of the comparative method has been borne out in the fields of anatomy, physiology, and medicine. Decisions and priorities regarding the choice of organisms for whole-genome sequencing will ultimately shape biology and influence the potential applications of the completed genome sequences. To assist in selecting future species of mammals for full or abridged genome sequence determinations, we should first examine the criteria and applications that various proposed species might offer.

A frequently raised argument in favor of comparative genome sequencing projects is their relevance to human biology. Criteria to select the most appropriate organisms for complete genome sequencing include: phylogenetic proximity to humans, whether the organism is used in research, the size of its genome, and the ease with which it can be genetically assessed through mutation (12). Accordingly, sequencing projects have begun for such species as mouse, rat, zebrafish, and pufferfish, with price tags in the $50 to $75 million range (6-8). Primate experts are lobbying to sequence the genome of our closest relative, the chimpanzee (9-11), which should provide clues to what makes us human. Proponents of a rhesus macaque genome sequence emphasize the value of this animal model for studying different human diseases and for testing new vaccines and drugs (13).

Although there is clear value to designating experimental model organisms, the criteria for selecting the next group of species to have their genomes sequenced are less obvious. Consideration of species not regularly thought of as "model organisms" may provide valuable insight into

the workings of modern genomes, as well as the dynamic evolutionary genetic processes that shaped and selected them. We suggest that evolutionary history and adaptive phylogenetic informativeness should be considered. Here is why.

Recently, independent molecular genetic analyses (with broad species sampling and multikilobase data sets of homologous gene segments) have produced a concordant view of the very earliest divergence events among the 18 modern orders of placental mammals (14-16). Four principal lineages or "clades" (labeled I to IV in the figure) have been obtained from the partitioning of placental mammal precursors. (This partitioning may have been a consequence of the Cretaceous supercontinental breakup, some 70 to 100 million years ago.) The mammal species already scheduled for sequencing (human, mouse, rat) are nested in one of the four principal clades (III in the figure). Thus, three of four surviving placental mammal lineages are not represented by current sequencing plans.



Order among the mammals. Phylogenetic relationships among modern orders of placental mammals (**left**) and candidate species for large-scale genome sequencing (**right**). Relevant considerations for species selection include biomedical relevance (M); economic importance (E); derived or specialized morphology (DM); ancestral or primitive morphology (AM); derived rearranged genome versus ancestral genome organization (DG) (22); ancestral or primitive genome organization (AG) (22); questionable or unknown genome organization (QG); phylogenetically close to human (CH); phylogenetically far from human (FH).

There is little dispute that the mouse is unrivaled for its power as a model for genetic analysis, and the same is true for the rat's value in medical physiology (6-8). There are cogent reasons, however, for considering additional mammals in the future. The mouse and rat are characterized by high rates of nucleotide substitution relative to other mammals (17), and the genomes of both species are rearranged extensively relative to both the human genome and the imputed genome of a primitive mammal (4). Both of these features result in a derived murine genome organization that may not mirror that of other mammals or, for that matter, of other rodents (4). This could be remedied by selecting mammal species with more ancestral (general) rather than more derived (specialized) genome characteristics.

After the primates and rodents (just 2 of 18 placental orders), which species would best promote our understanding of genome structure and function? Livestock and pets are important in the worldwide economy and also provide valuable models for scores of human genetic diseases, many of which are not represented in rodent strains (4). A Chinese-Danish consortium recently announced plans to sequence the pig genome (18). With similar reasoning, the cattle genome should be another top priority. Two pet species from the order Carnivora, cat and dog, are at an advanced stage of genomic resource development and are valuable veterinary models for hundreds of human hereditary and infectious diseases. These two species are starkly different when it comes to the organization of their genomes: The cat genome is highly conserved, whereas the dog genome is highly rearranged (4). An additional advantage is that livestock and pets--belonging to mammalian clade IV, the sister group to clade III containing rodents and primates--are close evolutionary reference (outgroup) species (see the figure). Beyond these representatives from clades III and IV, the next choices become more arbitrary. Sequencing one representative from each mammalian order would provide broad sampling, although a more economical strategy for selecting species might offer greater immediate benefits from the standpoints of both phylogenetic depth and genome characteristics.

At a minimum, at least one representative from each of the basal mammalian clades II-Xenarthra (sloths, anteaters, armadillos) and I-Afrotheria (which includes elephants, sea cows, aardvarks, and elephant shrews), as well as a marsupial and a monotreme species, should be considered for large-scale genome analysis (see the figure). It may be useful to choose mammals with primitive body plans and life strategies (such as shrews, tree shrews, and tenrecs), because these species seem to have retained a genome that specifies the primitive or ancestral developmental program for all mammals. Alternatively, investigation of species with diverse, derived morphologies (such as dolphins, bats, and elephants) may provide insight into the genomic basis of macroevolutionary changes in mammalian development. The same advice would hold for the investigation of genome organization patterns, where conserved primitive genomes (as illustrated by those of cat, pig, dolphin, and shrew) might reveal more general features in contrast to more derived reshuffled genomes (such as those of mouse, dog, bear, and gibbon) (4).

Small genome size--the rationale for sequencing the pufferfish (400 million base pairs)--might be a reason to select mammals with smaller genomes such as bats (~1.7 million base pairs). Comparing a small genome to the human, rat, and mouse genomes would aid in identifying critical noncoding genome regions and would increase our understanding of repetitive elements (SINES, LINES, and long terminal repeats) that constitute 50% of the human sequence. Another possible consideration is species diversity. Some orders retain many species (rodents, 2052; bats, 977), whereas others have far fewer (Proboscidea, 2; Tubulidentata, 1) (see the figure). As has been suggested for species conservation management (19, 20), species richness and poorness within extant orders should not be overlooked. Financial constraints will limit the spectrum of whole-genome sequencing projects to those species of extraordinary value to biologists. Indeed, whole-genome

sequencing may be phylogenetically uninformative for much of the noncoding regions that cannot be confidently aligned between mammalian orders. Because less than 2% of the human genome is coding sequence (2, 3), this may become a serious concern. For selected species, an abridged genome sequencing strategy--involving the targeted sequencing of particular genomic segments aligned with human-mouse-rat index sequence maps--should be considered. Alternatively, full-length cDNA sequencing projects for candidate species would provide more economical but still expansive comparative genomics databases (21). These cDNA gene collections, currently under way for mice and humans, will greatly facilitate functional evolutionary analyses using microarray technology. Such abridged genome sequence resources, when interpreted in the greater context of aligned finished sequences from index species, would offer a rich comparative baseline for developmental, physiologic, and functional genomic inference.

We have raised several criteria that should be considered, or at least acknowledged, as part of the choice to undertake large-scale genome sequencing. The considerations include: (i) phylogeny; (ii) relevance to understanding human biology and medicine, clearly appropriate to lab animal models (mouse, rat, cat, dog) and certain primates; (iii) economic importance, as exemplified by domesticated agricultural species (cow, pig, sheep); (iv) genomic characteristics, including genome size, chromosome conservation, and other conserved features; (v) recognition of developmentally extreme morphological specialization (derived) and conserved primitive (general) features; and (vi) species diversity among mammalian orders. These factors should be evaluated preferably by specialists for each lineage who can estimate the projected value expected from selection of a species based on scientific rather than subjective preferences.

There is no correct answer to species selection, only better or poorer choices. Evolutionary genomic considerations may provide some guidance and promote dialogue so that the human genome project will benefit maximally from comparative genetic principles, by selecting species that offer the greatest potential for a fuller understanding of the human genome and the complex evolutionary processes that created it.

References

1. J. Rettie, *Coronet* **29**, 21 (1951).
2. J. C. Venter *et al*., *Science* **291**, 1304 (2001).
3. E. S. Lander *et al*., *Nature* **409**, 860 (2001) [Medline].
4. S. J. O'Brien *et al*., *Science* **286**, 458 (1999).
5. G. Rubin *et al*., *Nature* **409**, 820 (2001) [Medline].
6. E. Pennisi, *Science* **287**, 1179 (2000) 1179.
7. J. Nadeau *et al*., *Science* **291**, 1251 (2001) .
8. E. Pennisi, *Science* **291**, 1204 (2001).
9. A. Gibbons, *Science* **289**, 1267 (2000).
10. E. H. McConkey *et al*., *Science* **289**, 1295 (2000).
11. S. Paabo, *Science* **291**, 1219 (2001).
12. F. S. Collins *et al*., *Science* **282**, 682 (1998).
13. J. L. VandeBerg *et al*., *Science* **290**, 1504 (2000).
14. W. H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997) [publisher's information].
15. W. J. Murphy *et al*., *Nature* **409**, 614 (2001) [Medline].
16. O. Madsen *et al*., *Nature* **409**, 610 (2001) [Medline].
17. E. Eizirik *et al*., *J. Hered*. **92**, 212 (2001) [Medline].
18. V. G. Cheung *et al*., *Nature* **409**, 953 (2001) [Medline].
19. R. I. Vane-Wright *et al*., *Biol. Conserv*. **55**, 235 (1991).
20. R. M. May, *Nature* **347**, 129 (1990).
21. R. L. Strausberg *et al*., *Science* **286**, 455 (1999).
22. W. J. Murphy *et al*., *Genome Biol.*, in press.

The authors are in the Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA. E-mail: obrien@ncifcrf.gov

This article has been cited by other articles:

- Premzl, M., Gready, J. E., Jermiin, L. S., Simonic, T., Marshall Graves, J. A. (2004). Evolution of Vertebrate Genes Related to Prion and Shadoo Proteins--Clues from Comparative Genomic Analysis. *Mol Biol Evol* 21: 2210-2231 [Abstract] [Full Text]
- Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K., de Jong, P., Nierman, W. C., Strausberg, R. L., Fraser, C. M. (2004). Human, Mouse, and Rat Genome Large-Scale Rearrangements: Stability Versus Speciation. *Genome Res.* 14: 1851-1860 [Abstract] [Full Text]
- Larkin, D. M., Everts-van der Wind, A., Rebeiz, M., Schweitzer, P. A., Bachman, S., Green, C., Wright, C. L., Campos, E. J., Benson, L. D., Edwards, J., Liu, L., Osoegawa, K., Womack, J. E., de Jong, P. J., Lewin, H. A. (2003). A Cattle-Human Comparative Map Built with Cattle BAC-Ends and Human Genome Sequence. *Genome Res.* 13: 1966-1972 [Abstract] [Full Text]
- McCue, L. A., Thompson, W., Carmack, C. S., Lawrence, C. E. (2002). Factors Influencing the Identification of Transcription Factor Binding Sites by Cross-Species Comparison. *Genome Res.* 12: 1523-1532 [Abstract] [Full Text]
- Murphy, W. J., Page, J. E., Smith, C. Jr., Desrosiers, R. C., O'Brien, S. J. (2001). A Radiation Hybrid Mapping Panel for the Rhesus Macaque. *J Hered* 92: 516-519 [Abstract] [Full Text]
- Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W., Springer, M. S. (2001). Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics. *Science* 294: 2348-2351 [Abstract] [Full Text]
- Daly, D. C., Cameron, K. M., Stevenson, D. W. (2001). Plant Systematics in the Age of Genomics. *Plant Physiol.* 127: 1328-1333 [Full Text]

Volume 292, Number 5525, Issue of 22 Jun 2001, pp. 2264-2266.